OPTIMIZING SPEECH/NON-SPEECH CLASSIFIER DESIGN USING ADABOOST

Oh-Wook Kwon and Te-Won Lee

Institute for Neural Computation, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0523
owkwon@ucsd.edu, tewon@ucsd.edu

ABSTRACT

We propose a new method to design speech/non-speech classifiers for voice activity detection and robust endpoint detection using the adaptive boosting (AdaBoost) algorithm. The method uses a combination of simple base classifiers through the AdaBoost algorithm and a set of optimized speech features combined with spectral subtraction. The key benefits of this method are the simple implementation and low computational complexity. The AdaBoost classifier combined with spectral subtraction significantly improved the receiver operating characteristic curves of the G.729 voice activity detector. For speech recognition purpose, the method reduced 20–50% of miss errors for the same false alarm rate by using additional band pass energy and spectral distortion based on mel frequency cepstral coefficients.

1. INTRODUCTION

As speech recognition technologies are recently applied to portable devices in realistic noisy environments, robust speech detection has become one of the most critical components for speech recognition systems. Speech detection or endpoint detection has turned out to significantly influence word accuracy in case of cellular phones in a noisy automobile environment [1][2]. Conventional endpoint detection algorithms based on energy and zero crossing rate (ZCR) cannot handle noisy speech signals properly in mobile communications. Hence, additional features such as high-pass/low-pass energies, linear prediction coding (LPC) residual and autocorrelation of LPC residual information have been explored to improve robustness and accuracy. For noisy speech detection, speech enhancement stages are often adopted before speech/non-speech classification to reduce noise and therefore a new endpoint detector should be designed for enhanced speech signals. A necessary and strong constraint for speech/non-speech classification algorithms is that they should have as low computational complexity as possible to reduce computational burden on the entire speech recognition system.

To satisfy the constraint, we use an adaptive boosting (AdaBoost) algorithm [5] [6] to design speech/non-speech classifiers. The speech/non-speech classifier designed with the AdaBoost algorithm combines very simple base classifiers to achieve the accuracy similar to the manually-optimized G.729 voice activity detector (VAD) [7] with comparable computational complexity. We achieved classification performance superior to the G.729 VAD by adopting and weighting new features for spectral subtracted signals. Further analysis of the learned weights for the base classifiers revealed the contribution of each feature component. The proposed method was evaluated using the Aurora database for voice activity detection and robust speech detection.

2. SPEECH/NON-SPEECH CLASSIFICATION USING THE ADABOOST ALGORITHM

2.1. Voice Activity Detector

The G.729 VAD uses the following features for speech/non-speech classification in the first stage [7]: (1) instantaneous full-band log energy, (2) low-band log energy difference, (3) full-band log energy difference, (4) spectral distortion measured by line spectral frequencies, and (5) zero-crossing rate difference. Each difference feature is obtained by the difference between the instantaneous parameter and the running average of the background noise. For each frame, the initial VAD decision was made by using a speech/non-speech classifier with 14 hyperplanes whose parameters were determined by visual inspection over a large database. For simplicity of the design, each hyperplane uses only two features.

In our design we keep the simplicity of the speech/non-speech classifier of the G.729 Annex B while we train the hyperplanes in a principled and automatic manner using the AdaBoost algorithm. All differential features were normalized to have zero mean and unit variance along each axis. We used a perceptron as the base classifier with the sigmoidal activation function $f(\mathbf{x}) = \tanh(\gamma \mathbf{x})$ where $\mathbf{x}$ is the feature vector and $\gamma = 4$ is used to control the range of boundary regions. The sign of the base classifier output $h_t(\mathbf{x})$ is the predicted label and the magnitude denotes a measure of confidence. We use base classifiers with linear

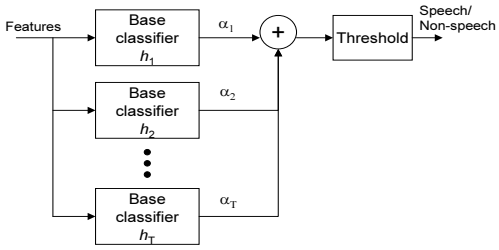

Fig. 1. Speech/Non-speech classifier using the AdaBoost algorithm.

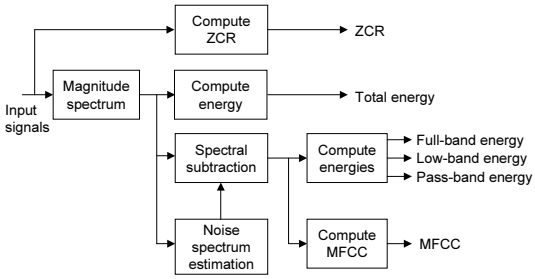

Fig. 2. Feature extraction in combination with spectral subtraction.

decision boundaries due to fast and simple learning procedures. For $h_t(\mathbf{x})$, we chose to use a decision stump where only a single feature is used in each perceptron for its simplicity. Experimental results showed that performance difference between the decision stump and the general perceptron was not significant. The final classifier is given by $H(\mathbf{x}) = \text{sign}(\sum_{t=1}^{T} \alpha_t(h_t(\mathbf{x}) + \delta))$ where α_t is the weight for the t-th base classifier and δ is a parameter to control the hit and false alarm rates.

Fig. 1 shows the block diagram of the speech/non-speech classifier based on adaptive boosting [5]. When only one feature is used for each classifier, it partitions the feature space into vertical or horizontal decision boundaries. The decision boundary of the final classifier become non-linear because the signum function is used to combine the base classifiers.

For spectral subtraction, the noise spectrum is estimated from the input signals and subtracted from the magnitude spectrum of input signals. First the smoothed spectrum was obtained by filtering in the spectral direction as well as in the temporal direction. Then the noise spectrum was estimated by tracking the minimum statistics of the magnitude spectrum [4], where the minimum of each frequency bin within the time window of 1 second was regarded as a noise component. The over-subtraction technique was also applied to reduce musical noise. This method does not require any other assumption on input speech utterances and can be used continuously without reinitialization.

2.2. For Robust Endpoint Detection

The G.729 VAD is targeted for speech signals with rather low level of noise signals and its performance degrades as the signal-to-noise ratio (SNR) goes down to about 5 dB. Therefore a speech enhancement block was applied before feature extraction. In addition, it is advantageous to use mel-frequency cepstral coefficient (MFCC)-based features so that we can combine feature extraction and spectral subtraction to share the required computation with a speech recognizer.

Features for speech/non-speech classification are extracted as shown in Fig. 2: (1) full-band speech log energy difference, (2) low-band speech log energy difference (0 – 1.0 kHz), (3) pass-band speech log energy difference (1.0 – 2 kHz), (4) spectral distortion measured by MFCC, (5) zero crossing rate difference, and (6) instantaneous total log energy. We note that in this case the band energies are for spectral-subtracted signals. The low-pass log energy and band-pass log energy are useful to reject high-frequency noise (e.g., drill noise) and low-frequency noise (car noise) [3]. The MFCC-based distortion measure can be easily computed from the feature extraction of the standard speech recognizers.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

3.1. Speech Database

The Aurora speech database [2] was used to evaluate the performance of the proposed speech detector. The speech/non-speech classifiers were trained with speech data in all environments and all SNR levels available in the database. For the test set, the speech data with the same noise environments were used in the same range of SNRs. Label information was obtained by using a Viterbi aligner obtained from multi-style training, manually corrected by viewing the spectrogram and used to train the classifiers. We used randomly-sampled 12000 frames as the training data set and another 3000 frames as the validation data set. Speech data with babble noise in different SNR conditions were used as the test data set.

3.2. Voice Activity Detection

We did not use the hang-over scheme [7] to compare only the speech/non-speech classification performance. The number of base classifiers used in the test was decided to give the minimum error rate for the validation test set. In this case the number of the base classifiers used in the test was 95. The performance points of the G.729 VAD were located near or on the receiver operating characteristic (ROC) curves of our proposed method. However, the AdaBoost-based classification has the advantage that it can provide a

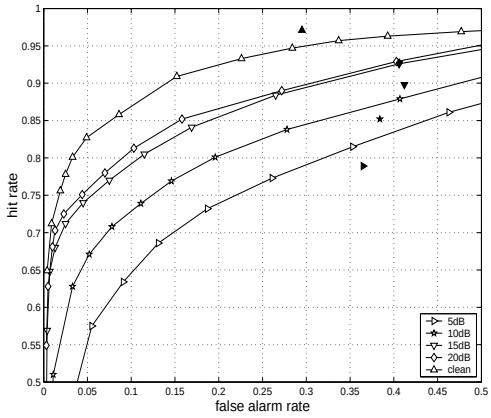

Fig. 3. ROC curves of voice activity detection of spectral-subtracted signals. The filled symbols denote the performance of the G.729 VAD without any additional processing, where the performance is shown as a point for each SNR condition.

flexible trade-off between the hit rate and the false alarm rate by controlling only one parameter δ depending on applications.

We trained the AdaBoost classifier using the speech signals enhanced by spectral subtraction. To obtain the ROC curves for spectral subtracted signals shown in Fig. 3, we normalized α_t to have unity sum of absolute value of α_t and varied δ discretely from -0.3 to 0.5 with a step size of 0.05. Each symbol in the ROC curve denotes the condition with a certain control parameter. The filled symbols denote the case of the G.729 VAD without any additional processing, where the performance is shown as a point for each SNR condition because the G.729 VAD has a fixed parameter. Although its performance degraded slightly in the clean speech case, the hit rate at the same false alarm rate was improved in the noisy cases. One important advantage in using the AdaBoost-based classifier is that we can automatically obtain a simple classifier with performance comparable to the manually optimized classifier.

The performance of the G.729 with spectral subtraction was not plotted because the operating points were mostly out of the current plot range. This is due to the changes in the speech signal characteristic induced by the spectral subtraction algorithm and the G.729 VAD cannot adapt to the distorted signals.

3.3. Robust Endpoint Detection

For robust endpoint detection, we attempted to use better features derived from the feature extraction module for speech recognition as described in Section 2.2. As shown in Fig. 4, we analyzed the learned weight values and found that the low- and pass-band energies largely contribute to the

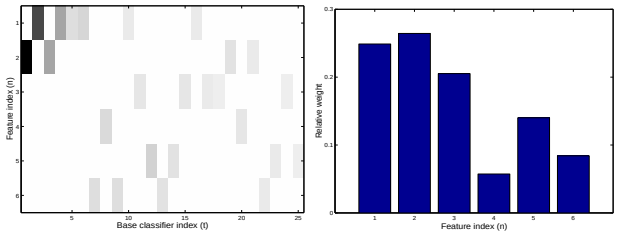

Fig. 4. Feature weights for the MFCC case for the first 25 base classifiers (left) and relative weight of each feature component averaged all base classifiers (right). In the left figure, the darkness of a square denotes the magnitude of the corresponding weight.

performance of the classifier. In particular, the low-band speech log energy has a relatively large weight and the spectral distortion feature has a lower weight than the G.729 VAD because spectral subtraction introduced spectral distortion. However, the MFCC-based spectral distortion feature did not have a large weight because the spectral subtraction caused nonlinear distortion on the spectrum of the input signals. To the contrary, the instantaneous full-band log energy and the low-band energy were the most important two features in case of the G.729-based features.

Fig. 5 shows the ROC curves in the babble noise environments using spectral subtraction and the new features. In every SNR condition the AdaBoost-based speech/non-speech classifier yielded improved performance. For the same false alarm rate condition as in the G.729 VAD, the miss rate ($= 1 - \text{hit rate}$) decreased by 20–50 percent. This improvement mainly results from spectral subtraction and the proper design of the classifier as the features are changed.

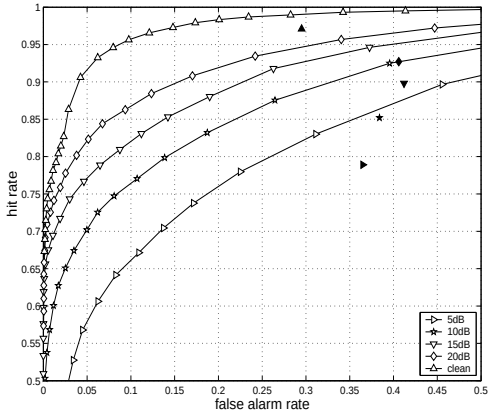

Fig. 5. ROC curves in case of new features. Spectral subtraction was applied before feature extraction.

We compared the performance of the G.729 VAD with 10 dB SNR as shown in Fig. 6: the AdaBoost classifier with

G.729 features, the AdaBoost classifier with G.729 features and spectral subtraction, and the AdaBoost classifier with the new features and spectral subtraction. Fig. 6 shows that the AdaBoost-based classifier yield the same level of performance similar to the manually-optimized classifier. The AdaBoost classifier was systematically designed by multiple decision stumps, which requires a small number of computations. For better performance we changed the feature and improved the performance with minimal increase of computational load.

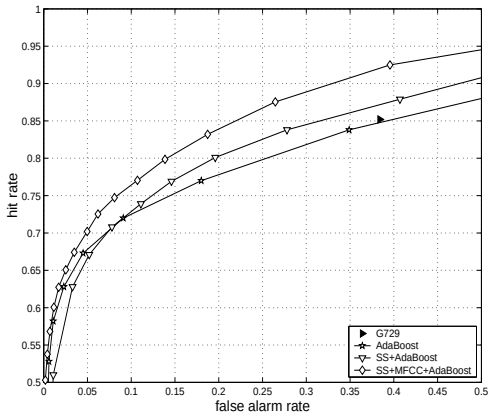

Fig. 6. Performance comparison of the G.729 VAD in 10 dB SNR condition: the original G.729 VAD ('G729'), the AdaBoost with G.729 features ('AdaBoost') the AdaBoost with G.729 features and spectral subtraction ('SS + AdaBoost'), and the AdaBoost with the new features and spectral subtraction ('SS + MFCC + AdaBoost').

The experimental results, though not given in this paper, showed that the AdaBoost classifier also yields similar performance in car noise environments. Due to the nature of the car noise, the false alarm rates were relatively small compared to the babble noise case.

3.4. Considerations on Computational Complexity

The decision stump-based speech/non-speech classifier requires 1 addition of the bias term, 1 table lookup for $\tanh(.)$ and 1 multiplication by the weight for each base classifier. On the contrary, the speech/non-speech classifier for the G.729 VAD uses 14 hyperplanes with 2 features involved for each hyperplane and the final decision is made through the sequential test of the hyperplanes. Each hyperplane needs one multiplication, one addition and one comparison. In both cases, the required computation is linearly proportional to the number of base classifiers or hyperplanes, which can be tuned for specific applications.

4. CONCLUSIONS

We proposed a new method to design a speech/non-speech classifier based on the AdaBoost algorithm. Our experimental results indicate that a nearly optimal classifier can be designed automatically with computational complexity comparable to the G.729 VAD. The contribution of each feature and the effects of spectral subtraction were also investigated. The AdaBoost classifier with spectral subtraction significantly improved the ROC curves of the G.729 VAD. For speech recognition purposes, we suggested a new efficient endpoint detection method using different kinds of features including estimated speech band energies and MFCC-based spectral distortion. When spectral subtraction is used, the low-band speech log energy has a relatively large weight and the spectral distortion feature has a small weight. The proposed method to design a speech/non-speech classifier is highly practical and the classifier outperforms the current industrial VAD algorithm.

5. REFERENCES

[1] H.K. Kim, R.C. Cox, "Evaluation of robust speech recognition algorithms for distributed speech recognition in a noisy automobile environment," In *Proc. ICSLP 2002*, pp. 233-236, Sept. 2002.

[2] C.-P. Chen, K. Filali, J.F. Bilmes, "Frontend post-processing and backend model enhancement on the Aurora 2.0/3.0 databases," In *Proc. ICSLP 2002*, pp. 241-244, Sept. 2002.

[3] M. Marzinzik, B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," In *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 2, pp. 109-110, Feb. 2002.

[4] R. Martin, "Spectral subtraction based on minimum statistics," In *Signal Processing VII, Theories and Applications. Proc. EUSIPCO-94*, pp. 1182-1185, 1994.

[5] R.E. Schapire, Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, 37(3), pp. 297-336, 1999.

[6] Y. Freund, R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, 55(1), pp. 119-139, 1997.

[7] A. Benyassine, E. Shlomot, and H.-Y. Su, "ITU-T Recommendation G.729 Annex B: A silent compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, pp. 64-73, Sept. 1997.